Extreme-Scale UQ for Bayesian Inverse Problems Governed by PDEs

Tan Bui-Thanh*, Carsten Burstedde*†, Omar Ghattas*‡,
James Martin*, Georg Stadler*, Lucas C. Wilcox*§

*Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, Austin, TX
†Now at Institute for Numerical Simulation, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany
‡Jackson School of Geosciences, and Department of Mechanical Engineering, The University of Texas at Austin, Austin, TX
§Now at Department of Applied Mathematics, Naval Postgraduate School, Monterey, CA

***Abstract*—Quantifying uncertainties in large-scale simulations has emerged as the central challenge facing CS&E. When the simulations require supercomputers, and uncertain parameter dimensions are large, conventional UQ methods fail. Here we address uncertainty quantification for large-scale inverse problems in a Bayesian inference framework: given data and model uncertainties, find the pdf describing parameter uncertainties. To overcome the curse of dimensionality of conventional methods, we exploit the fact that the data are typically informative about low-dimensional manifolds of parameter space to construct low rank approximations of the covariance matrix of the posterior pdf via a matrix-free randomized method. We obtain a method that scales independently of the forward problem dimension, the uncertain parameter dimension, the data dimension, and the number of cores. We apply the method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with over one million uncertain earth model parameters, 630 million wave propagation unknowns, on up to 262K cores, for which we obtain a factor of over 2000 reduction in problem dimension. This makes UQ tractable for the inverse problem.**

I. Introduction

Perhaps the central challenge facing the field of computational science and engineering today is: *how do we quantify uncertainties in the predictions of our large-scale simulations, given limitations in observational data, computational resources, and our understanding of physical processes* [1]. For many societal grand challenges, the "single point" deterministic predictions delivered by most contemporary large-scale simulations of complex systems are just a first step: to be of value for decision-making (design, control, allocation of resources, policy-making, etc.), they must be accompanied by the degree of confidence we have in the predictions. Examples of problems for which large-scale simulations are playing an increasingly important role for decision-making include: mitigation of global climate change, natural hazard forecasts; siting of nuclear waste repositories, monitoring of subsurface contaminants, control of carbon sequestration processes, management of the nuclear fuel cycle, design of new nano-structured materials and energy storage systems, and patient-specific planning of surgical procedures, to name a few.

Unfortunately, when the simulations (here assumed without loss of generality to comprise PDEs) are expensive, and the uncertain parameter dimension is large (or even just moderate), conventional uncertainty quantification methods fail dramatically. Here we address uncertainty quantification (UQ) in large-scale inverse problems governed by PDEs. This is the crucial step in UQ: before we can propagate parameter uncertainties forward through a model, we must first infer them from observational data and from the (PDE) model that maps parameters to observables; i.e., we must solve the inverse problem. We adopt the Bayesian inference framework [2], [3]: given observational data and their uncertainty, the governing forward PDEs and their uncertainty, and a prior probability distribution describing prior uncertainty in the parameters, find the posterior probability distribution over the parameters, which is seen as the solution of the inverse problem. The grand challenge in solving statistical inverse problems is in computing statistics of the posterior probability density function (pdf), which is a surface in high dimensions. This is notoriously challenging for statistical inverse problems governed by expensive forward models (as in our target case of global seismic wave propagation) and high-dimensional parameter spaces (as in our case of inferring a heterogeneous parameter field). The difficulty stems from the fact that evaluation of the probability of each point in parameter space requires solution of the forward problem (which may tax contemporary supercomputers), and many such evaluations (millions or more) are required to adequately sample the posterior density in high dimensions by conventional Markov-chain Monte Carlo (MCMC) methods. Thus, UQ for the large-scale inverse problems becomes intractable.

The approach we take is based on a linearization of the parameter-to-observable map, which yields a local Gaussian approximation of the posterior. The mean and covariance of this Gaussian can be found from an appropriately weighted regularized nonlinear least squares optimization problem, which is known as the *maximum a posteriori* (MAP) point. The solution of this optimization problem provides the mean, and the inverse of the Hessian matrix of the least squares function (evaluated at the MAP point) gives the covariance matrix. Unfortunately, the most efficient algorithms available for direct computation of the (nominally dense) Hessian are prohibitive, requiring as many forward PDE-like solves as there are uncertain parameters, which can number in the

SC12, November 10-16, 2012, Salt Lake City, Utah, USA
978-1-4673-0806-9/12/$31.00 ©2012 IEEE

millions or more when the parameter represents a field (e.g, initial condition, heterogeneous material coefficient, source term).

The key insight to overcoming this barrier is that the data are typically informative about a low dimensional manifold of the parameter space [4]—that is, the Hessian of the data-misfit term in the least squares function is sparse with respect to some basis. We exploit this fact to construct a low rank approximation of the data-misfit Hessian and the resulting posterior covariance matrix using a parallel, matrix-free randomized algorithm, which requires a *dimension-independent* number of forward PDE solves and associated adjoint PDE solves (the latter resemble the forward PDEs in reverse time). UQ thus reduces to solving a fixed (and often small, relative to the parameter dimension) number of PDEs. When scalable solvers are available for the forward PDEs, *the entire process of quantifying uncertainties in the solution of the inverse problem is scalable with respect to PDE state variable dimension, uncertain parameter dimension, observational data dimension, and number of processor cores.* We apply this method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with 1.067 million parameters and 630 million wave propagation spatial unknowns over 2400 time steps, on up to 262,144 Jaguar cores. The example demonstrates independence of parameter dimension and a factor of over 2000 reduction in problem dimension. This UQ computation is orders of magnitude larger than any attempted before on a large-scale forward problem.

We recently presented a finite-dimensional version of our method (in which Lanczos iterations are used to build the low rank approximation of the Hessian) and applied it to a 1D inverse problem in moderate dimensions [5]. We have also recently described the extension to infinite-dimensional inverse problems (so-called because the inversion parameters represent a field) in the framework of [6], in which we discuss mathematically subtle yet critical issues related to the proper choice of prior and to discretizations that assure convergence to the correct infinite-dimensional quantities [7]. In this, our Bell Prize submission in the Scalable Algorithms category, we extend the method to extreme-scale Bayesian inverse problems, employing a randomized parallel matrix-free low rank approximation method, instead of Lanczos. The randomized method yields a low rank approximation with controllably high probability, and is asynchronous, more robust, more fault tolerant, and provides better cache performance. In the following sections, we provide an overview of the Bayesian formulation of inverse problems (§II), describe how the mean and covariance of the posterior pdf can be approximated from the solution of a regularized weighted nonlinear least-squares problem (§III and §IV), present our algorithm for parallel low rank-based covariance approximation (§V), assert the scalability of the overall UQ method (§VI), apply our method to the Bayesian solution of a very large scale inverse problem in 3D global seismic wave propagation (§VII), and draw conclusions (§VIII).

II. Bayesian Formulation of Inverse Problems

In the Bayesian approach, we state the inverse problem as a problem of *statistical inference* over the space of uncertain parameters, which are to be inferred from the data and a PDE model. The resulting solution to the statistical inverse problem is a posterior distribution that assigns to any candidate set of parameter fields our belief (expressed as a probability) that a member of this candidate set is the "true" parameter field that gave rise to the observed data. When discretized, this problem of infinite dimensional inference gives rise naturally to a large scale problem of inference over the discrete parameter space $\boldsymbol{x} \in \mathbb{R}^n$, corresponding to degrees of freedom in the parameter field mesh. While the presentation in this paper is limited to the finite dimensional approximation to the infinite dimensional measure, the discretization process is performed rigorously following [6], [7], and the numerical evidence indicates that we converge to the correct infinite dimensional distribution.

The posterior probability distribution combines the prior pdf $\pi_{\text{prior}}(\boldsymbol{x})$ over the parameter space, which encodes any knowledge or assumptions about the parameter space that we may wish to impose before the data are considered, with a likelihood pdf $\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x})$, which explicitly represents the probability that a given set of parameters $\boldsymbol{x}$ might give rise to the observed data $\boldsymbol{y}_{\text{obs}} \in \mathbb{R}^m$. Bayes' Theorem then explicitly computes the posterior pdf as

$$\pi_{\text{post}}(\boldsymbol{x}|\boldsymbol{y}_{\text{obs}}) \propto \pi_{\text{prior}}(\boldsymbol{x})\pi_{\text{like}}(\boldsymbol{y}_{\text{obs}}|\boldsymbol{x}).$$

We choose the prior distribution to be Gaussian, with a covariance operator defined by the square of the inverse of an elliptic PDE operator. This choice yields several benefits. First, it enables implicit representation of the prior covariance operator as (the inverse of) a sparse operator, as opposed to traditional approaches that either store a dense covariance matrix or its approximation by principle vectors. Second, since the covariance operator is never needed explicitly—only its action on a vector is required— we are able to capitalize on fast $O(n)$ parallel elliptic solvers (in this paper, algebraic multigrid) to form this action via two elliptic solves. Third, the action of the symmetric square root factorization of the prior covariance is available explicitly (via one elliptic solve instead of two). Finally, this choice of covariance is useful for technical reasons, as it guarantees that samples from the prior distribution will be continuous.

The difference between the observables predicted by the model and the actual observations $\boldsymbol{y}_{\text{obs}}$ is due to both measurement and model errors, and is represented by the i.i.d. Gaussian random variable "noise" vector $\boldsymbol{e}$,

$$\boldsymbol{e} = \boldsymbol{y}_{\text{obs}} - \boldsymbol{f}(\boldsymbol{x}),$$

where $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$ is the (generally nonlinear) operator mapping model parameters to output observables. Then the pdf's for the prior and noise can be written in the form

$$\pi_{\text{prior}}(\boldsymbol{x}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_{\text{prior}})^T\boldsymbol{\Gamma}_{\text{prior}}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_{\text{prior}})\right),$$

and

$$\pi_{\text{noise}}(\boldsymbol{e}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{e}-\bar{\boldsymbol{e}})^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}(\boldsymbol{e}-\bar{\boldsymbol{e}})\right),$$

respectively, where $\bar{\boldsymbol{x}}_{\text{prior}}$ is the mean of the prior distribution, $\bar{\boldsymbol{e}}$ is the mean of the Gaussian noise, $\boldsymbol{\Gamma}_{\text{prior}} \in \mathbb{R}^{n\times n}$ is the covariance matrix for the prior, and $\boldsymbol{\Gamma}_{\text{noise}} \in \mathbb{R}^{m\times m}$ is the covariance matrix for the noise. Restating Bayes' theorem with these Gaussian pdf's, we find that the statistical solution of the inverse problem, $\pi_{\text{post}}(\boldsymbol{x})$, is given by

$$\pi_{\text{post}}(\boldsymbol{x}) \propto \exp\Big(-\frac{1}{2}\|\boldsymbol{x}-\bar{\boldsymbol{x}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}} - \frac{1}{2}\|\boldsymbol{y}_{\text{obs}}-\boldsymbol{f}(\boldsymbol{x})-\bar{\boldsymbol{e}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}}\Big), \quad (1)$$

Note that the seemingly simple expression $\boldsymbol{f}(\boldsymbol{x})$ belies the complexity of the underlying computations, which involve: (1) construction of the PDE model for given parameters $\boldsymbol{x}$; (2) solution of the governing PDE model to yield the output state variables; and (3) extraction of the observables from the states at the observation locations in space and time. In §VII, we provide expressions for the underlying mathematical operators for our target inverse seismic wave propagation problem, in which the parameters are wave speeds in the earth, the governing PDEs describe acoustic wave propagation, and the observations are of velocity waveforms at seismometer locations on earth's surface. In general, $\boldsymbol{f}(\boldsymbol{x})$ is nonlinear, even when the forward PDEs are linear in the state variables (as is the case for the seismic inverse problem), since the model parameters couple with the states nonlinearly in the forward PDEs.

As is clear from the expression (1), despite the choice of Gaussian prior and noise probability distributions, the posterior probability distribution need not be Gaussian, due to the nonlinearity of $\boldsymbol{f}(\boldsymbol{x})$. The non-Gaussianity of the posterior poses challenges for computing statistics of interest for typical large-scale inverse problems, since as mentioned in §I, π_{post} is often a surface in high dimensions (millions, in our target problem in §VII), and evaluating each point on this surface requires the solution of the forward PDEs (wave propagation equations with $O(10^9)$ unknowns, in the target problem). Numerical quadrature to compute the mean and covariance matrix, for example, is completely out of the question. The method of choice for computing statistics is Markov chain Monte Carlo (MCMC), which judiciously samples the posterior distribution, so that sample statistics can be computed. But the use of MCMC for large-scale inverse problems is still prohibitive for expensive forward problems and high dimensional parameter spaces, since even for modest numbers of parameters, the number of samples required can be in the millions. An alternative approach based on linearizing the parameter-to-observable map is discussed next.

III. Posterior mean approximation

The mean of the posterior distribution $\bar{\boldsymbol{x}}_{\text{post}}$ can be approximated by finding the point that maximizes the posterior pdf, i.e., the MAP point,

$$\bar{\boldsymbol{x}}_{\text{post}} \approx \boldsymbol{x}_{\text{MAP}} := \arg\max_{\boldsymbol{x}} \pi_{\text{post}}(\boldsymbol{x}).$$

This approximation is exact when the map from parameters to observables, $\boldsymbol{f}(\boldsymbol{x})$, is linear. Finding the MAP point is equivalent to minimizing the negative log of the posterior pdf, i.e.,

$$\bar{\boldsymbol{x}}_{\text{post}} \approx \arg\min_{\boldsymbol{x}} V(\boldsymbol{x}), \quad (2)$$

where

$$V(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{y}_{\text{obs}}-\boldsymbol{f}(\boldsymbol{x})-\bar{\boldsymbol{e}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{noise}}} + \frac{1}{2}\|\boldsymbol{x}-\bar{\boldsymbol{x}}_{\text{prior}}\|^2_{\boldsymbol{\Gamma}^{-1}_{\text{prior}}}. \quad (3)$$

Approximating the mean of the posterior distribution by finding the MAP point is thus equivalent to solving a regularized deterministic inverse problem, where $\boldsymbol{\Gamma}^{-1}_{\text{prior}}$ plays the role of the regularization operator, and $\boldsymbol{\Gamma}^{-1}_{\text{noise}}$ is a weighting for the data misfit term.

Here, we solve the nonlinear least squares optimization problem (2) with a parallel inexact Newton–conjugate gradient method. The method requires the computation of gradients and Hessian-vector products of $V(\boldsymbol{x})$ (for which expressions are provided in §VII in the context of the seismic inverse problem we target). Rather than provide a detailed description of the method here, we refer to our earlier work presented at SC2002 [8] and SC2003 [9] on parallel scalability of the method, as well as the recent work [10] that includes additional refinements. The main ingredients of the method are:

- inexact Hessian matrix-free Gauss-Newton-conjugate gradient (CG) minimization;
- preconditioning by $\boldsymbol{\Gamma}^{-1}_{\text{prior}}$, carried out by multigrid V-cycles on the underlying elliptic operators;
- Armijo-type backtracking line search globalization;
- computation of gradients of $V(\boldsymbol{x})$ and products of Hessians of $V(\boldsymbol{x})$ with vectors at each CG iteration expressed as solutions of forward and (backward-in-time) adjoint PDEs and their linearizations, all of which inherit the parallel scalability properties of the forward PDE solver;
- algorithmic checkpointing to implement the composition of forward-in-time forward PDE solutions and backward-in-time adjoint PDE solutions to form gradients without having to store the entire state variable time history; and
- parallel implementation of all components of the method, which are dominated by solution of forward and adjoint-PDEs and evaluation of inner product-like quantities to compose gradient and Hessian-vector quantities.

What can be said about parallel and algorithmic scalability of this method? Because the dominant components of the method can be expressed as solutions or evaluations of PDE-like systems, parallel scalability—that is, maintaining high parallel efficiency as the number of cores increases—is assured whenever a scalable solver for the underlying PDEs is available (which is the case for our target seismic wave propagation problem [11]). The remaining ingredient to obtain overall scalability is that the method exhibit algorithmic scalability,

that is with increasing problem size. This is indeed the case: for a wide class of nonlinear inverse problems, the outer Newton iterations and the inner CG iterations are independent of the mesh size (as is the case for our target inverse wave propagation problem, [10]). This is a consequence of the use of a Newton solver, of the compactness of the Hessian of the data misfit term (i.e., the first term on the right hand side of (3), as proven for the inverse wave propagation setting in [4]), and the choice of prior preconditioning so that the resulting preconditioned Hessian is a compact perturbation of the identity, for which CG exhibits mesh-independent iterations. Thus, solving the least squares optimization problem (2) to approximate the mean of the posterior distribution by the method outlined above exhibits both parallel and algorithmic—and thus overall—scalability.

As stated above, the focus of this paper is not on the computation of the posterior mean $\bar{\boldsymbol{x}}_{\text{post}}$, but on the significantly more challenging task of characterizing the uncertainty in the mean via computation of the posterior covariance matrix, $\boldsymbol{\Gamma}_{\text{post}} \in \mathbb{R}^{n\times n}$. Linearizing the parameter-to-observable map at the MAP point gives

$$\boldsymbol{f}(\boldsymbol{x}) \approx \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{x}_{\text{MAP}}) + \boldsymbol{f}(\boldsymbol{x}_{\text{MAP}}),$$

where $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ is the Jacobian matrix of $\boldsymbol{f}(\boldsymbol{x})$ evaluated at $\boldsymbol{x}_{\text{MAP}}$. Manipulation of (1) shows that $\boldsymbol{\Gamma}_{\text{post}}$ is given by the inverse of the Hessian matrix of the function $V(\boldsymbol{x})$ in (3), i.e.,

$$\boldsymbol{\Gamma}_{\text{post}} = \left(\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A} + \boldsymbol{\Gamma}_{\text{prior}}^{-1}\right)^{-1}. \tag{4}$$

In summary, under the assumptions of this section (additive Gaussian noise, Gaussian prior, and linearized parameter-to-observable map), solution of the Bayesian inverse problem is reduced to the characterization of the (Gaussian) posterior distribution $\mathcal{N}(\bar{\boldsymbol{x}}_{\text{MAP}}, \boldsymbol{\Gamma}_{\text{post}})$, where $\boldsymbol{\Gamma}_{\text{post}}$ is the inverse of the Hessian of $V(\boldsymbol{x})$ at $\boldsymbol{x}_{\text{MAP}}$.

The primary difficulty here is that the large parameter dimension n prevents any representation of the posterior covariance $\boldsymbol{\Gamma}_{\text{post}}$ as a dense operator. In particular, the Jacobian of the parameter-to-observable map, $\boldsymbol{A}$, is formally a dense matrix, and requires n forward PDE solves to construct. This is intractable when n is large and the PDEs are expensive, as in our case. However, a key feature of the operator $\boldsymbol{A}$ is that its action on a (parameter field-like) vector can be formed by solving a (linearized) forward PDE problem; similarly, the action of its transpose $\boldsymbol{A}^T$ on a (observation-like) vector can be formed by solving a (linearized) adjoint PDE. Explicit expressions for these operations will be given for our specific target inverse problem in §VII. In the next two sections, we present algorithms that exploit this property, as well as the spectral decay of the data misfit Hessian, to approximate the posterior covariance matrix with controlled accuracy at a cost that is independent of the parameter dimension.

IV. Posterior covariance approximation

For many ill-posed inverse problems, the Hessian matrix of the data misfit term in (3), defined as

$$\boldsymbol{H}_{\text{misfit}} \overset{\text{def}}{=} \boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}, \tag{5}$$

is a discretization of a compact operator, i.e., its eigenvalues collapse to zero. This can be understood intuitively, since only the modes of the parameter field that strongly influence the observations (through the linearized parameter-to-observable map $\boldsymbol{A}$) will be present in the dominant spectrum of (5). In many ill-posed inverse problems, observations are sparse compared to the parameter dimensions, and numerous modes of the parameter field (for example, highly oscillatory ones) will have negligible effect on the observables. The range space thus is effectively finite-dimensional even before discretization (and therefore independent of any mesh), and the eigenvalues decay, often rapidly, to zero. In this section, we exploit this low-rank structure to construct scalable algorithms to approximate the posterior covariance operator.

Rearranging the expression for $\boldsymbol{\Gamma}_{\text{post}}$ in (4) to factor out $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ gives

$$\boldsymbol{\Gamma}_{\text{post}} = \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I}\right)^{-1}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{6}$$

This factorization exposes the *prior-preconditioned Hessian of the data misfit*,

$$\tilde{\boldsymbol{H}}_{\text{misfit}} \overset{\text{def}}{=} \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{7}$$

In the next section we present a randomized algorithm to construct a low rank approximation of this matrix at a cost (in PDE solves) that is independent of the parameter dimension (compared to n PDE solves to construct the full matrix). In this section, we assume only that such a low rank construction is possible. Let λ_i and $\boldsymbol{v}_i$ be the eigenvalues and eigenvectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_i) \in \mathbb{R}^{n\times n}$ be the diagonal matrix of its eigenvalues, and define as $\boldsymbol{V} \in \mathbb{R}^{n\times n}$ the matrix whose columns are the eigenvectors $\boldsymbol{v}_i$ of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Then replace $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by its spectral decomposition:

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} + \boldsymbol{I}\right)^{-1} = (\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T + \boldsymbol{I})^{-1}. \tag{8}$$

When the eigenvalues of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ decay rapidly, we can extract a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ by retaining only the r largest eigenvalues and corresponding eigenvectors,

$$\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_{\text{prior}}^{1/2} \approx \boldsymbol{V}_r\boldsymbol{\Lambda}_r\boldsymbol{V}_r^T.$$

Here $\boldsymbol{V}_r \in \mathbb{R}^{n\times r}$ contains only the r eigenvectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ that correspond to the r largest eigenvalues, which are assembled in the diagonal matrix $\boldsymbol{\Lambda}_r = \text{diag}(\lambda_i) \in \mathbb{R}^{r\times r}$. To obtain the posterior covariance matrix, we employ the Sherman-Morrison-Woodbury formula to perform the inverse

in (6),

$$\left(\boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{A}^T\boldsymbol{\Gamma}_{\text{noise}}^{-1}\boldsymbol{A}\boldsymbol{\Gamma}_{\text{prior}}^{1/2}+\boldsymbol{I}\right)^{-1} = \boldsymbol{I}-\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T+\mathcal{O}\left(\sum_{i=r+1}^{n}\frac{\lambda_i}{\lambda_i+1}\right),$$

where $\boldsymbol{D}_r \stackrel{\text{def}}{=} \text{diag}(\lambda_i/(\lambda_i+1)) \in \mathbb{R}^{r\times r}$. The last term in the expression above shows the error due to truncation in terms of the discarded eigenvalues; this provides a criterion for truncating the spectrum, namely r is chosen such that λ_r is small relative to 1. With this low-rank approximation, the final expression for the approximate posterior covariance follows from (6),

$$\boldsymbol{\Gamma}_{\text{post}} \approx \boldsymbol{\Gamma}_{\text{prior}} - \boldsymbol{\Gamma}_{\text{prior}}^{1/2}\boldsymbol{V}_r\boldsymbol{D}_r\boldsymbol{V}_r^T\boldsymbol{\Gamma}_{\text{prior}}^{1/2}. \tag{9}$$

Note that (9) expresses the posterior uncertainty (in the form of a covariance matrix) as the prior uncertainty, less any information gained from the data, filtered through the prior.

V. A randomized algorithm for low-rank Hessian approximation

We now address the construction of the low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ that was invoked in the previous section. As argued above, the data inform only a limited number of modes of the parameter field, resulting in a data misfit Hessian matrix that admits a low rank representation. This is observed numerically (see Figure 1) and has recently been proven theoretically in several settings [4], [12]. Moreover, preconditioning with the prior operator as in (7) further filters out modes of the parameter space that are already well-determined from prior knowledge (i.e., a smoothing prior such as the one we employ here assigns low probability to highly oscillatory modes.)

We exploit this structure to construct a low rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ using randomized algorithms for approximate matrix decomposition [13], [14]. Their performance is comparable to Krylov methods (such as Lanczos) we employed previously [5], [15]. However, they have a significant edge over these deterministic methods for large-scale problems, since the required Hessian matrix-vector products are independent of each other, providing asynchronousity and fault tolerance. Before discussing these advantages, let us summarize the algorithm.

To approximate the spectral decomposition of $\tilde{\boldsymbol{H}}_{\text{misfit}} \in \mathbb{R}^{n\times n}$, we generate a random matrix $\boldsymbol{R} \in \mathbb{R}^{n\times r}$ (r is of the order of the numerical rank of $\tilde{\boldsymbol{H}}_{\text{misfit}}$, so in our case $r \ll n$) with i.i.d. Gaussian entries, and compute the product $\boldsymbol{Y} = \tilde{\boldsymbol{H}}_{\text{misfit}}\boldsymbol{R}$. Since each column vector in $\boldsymbol{R}$ is an independent random vector, the computation of $\boldsymbol{Y}$ decouples into r separate matrix-vector product with $\tilde{\boldsymbol{H}}_{\text{misfit}}$. As can be seen from (7), each matrix-vector product requires a pair of forward/adjoint PDE solves (to form actions of $\boldsymbol{A}$ and $\boldsymbol{A}^T$ on vectors), as well as a pair of elliptic operator solves (to form actions of $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ on vectors). The latter are much cheaper than the former, in the typical case when the PDE model governing the inverse problem is large scale.

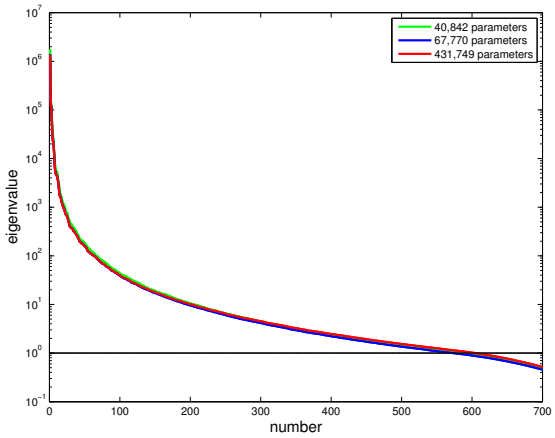

Fig. 1. Log-linear plot of the spectrum of prior-preconditioned data misfit Hessian ($\tilde{\boldsymbol{H}}_{\text{misfit}}$) for three successively finer parameter meshes of an inverse wave propagation problem [7]. The spectra lie on top of each other, indicating mesh independence (and therefore parameter-dimension independence) of the low rank approximation. The eigenvalues are truncated when they are small relative to 1, which in this case results in retaining between 0.2 and 2% of the spectrum.

Let $\boldsymbol{Q}$ be an orthonormal basis for $\boldsymbol{Y}$, which approximates the range space of $\tilde{\boldsymbol{H}}_{\text{misfit}}$. Following the "single-pass" approach of [13], we compute the approximation to $\tilde{\boldsymbol{H}}_{\text{misfit}}$ in the basis $\boldsymbol{Q}$:

$$\boldsymbol{B} \stackrel{\text{def}}{=} (\boldsymbol{Q}^T\boldsymbol{Y})(\boldsymbol{Q}^T\boldsymbol{R})^{-1} \approx \boldsymbol{Q}^T\tilde{\boldsymbol{H}}_{\text{misfit}}\boldsymbol{Q}. \tag{10}$$

Here $\boldsymbol{B}$, $\boldsymbol{Q}^T\boldsymbol{Y}$, and $(\boldsymbol{Q}^T\boldsymbol{R})^{-1}$ are all matrices of dimension r, which is much smaller than n, and thus we are able to decompose the symmetric matrix $\boldsymbol{B}$ as $\boldsymbol{Z}\boldsymbol{\Lambda}\boldsymbol{Z}^T$ using dense linear algebra. The dominant vectors of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ are then returned as $\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{Z}$, with eigenvalues on the diagonal of $\boldsymbol{\Lambda}$. Thus, we find the desired decomposition

$$\tilde{\boldsymbol{H}}_{\text{misfit}} \approx \boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}^T. \tag{11}$$

Finally, randomized methods also provide an estimate of the spectral norm of $\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T\tilde{\boldsymbol{H}}_{\text{misfit}}$, which bounds the error that we make in our low rank approximation. To be precise, the bound derived in [13] is

$$\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T)\tilde{\boldsymbol{H}}_{\text{misfit}}\right\| \le \alpha\sqrt{\frac{2}{\pi}}\max_{i=1,\ldots,r}\left\|(\boldsymbol{I}-\boldsymbol{Q}\boldsymbol{Q}^T)\boldsymbol{A}\boldsymbol{\omega}^{(i)}\right\|, \tag{12}$$

attained with probability of at least $1-\alpha^{-r}$, where $\boldsymbol{\omega}^{(i)}$ are vectors with i.i.d. standard normal entries.

To summarize, the construction of a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is dominated by its application to random vectors, which entails a pair of forward/adjoint PDE solves. The independence of these matrix-vector products from each other is of particular importance for problems in which the parameter-to-observable map $\boldsymbol{f}(\boldsymbol{x})$ has to be computed on large parallel supercomputers for the following reasons:

- Cache and memory efficiency: For parameter-to-observable maps that involve the solution of a PDE, the application of the Hessian to multiple vectors requires the solution of (linearized) forward/adjoint PDEs for multiple right-hand sides. Amortizing data movement over the multiple right-hand sides results in significantly greater memory and cache efficiency than can be obtained with sequential right-hand sides, as required by classical Krylov methods.
- Fault-tolerance: the construction of the low-rank matrix approximation is done as a post-processing step when a sufficient number of matrix-vector products is available. The asynchronous nature of the matrix-vector products provides greater fault tolerance (for example, the low rank approximation in §VII was computed using 10 different jobs with different run times and core counts ranging from 32K to 108K.

VI. Scalability of the UQ method

We now discuss the overall scalability of our UQ method to high-dimensional parameter spaces. First, we summarize the scalability of the construction of the low-rank-based approximate posterior covariance matrix in (9). As stated before, the linearized parameter-to-observable map $\boldsymbol{A}$ cannot be constructed explicitly, since it requires n linearized forward PDE solves. However, its action on a vector can be computed by solving a single linearized forward PDE, regardless of the number of parameters n and observations m. Similarly, the action of $\boldsymbol{A}^T$ on a vector can be computed by solving a linearized adjoint PDE. Moreover, the prior is usually much cheaper to apply than the forward or adjoint PDE solution (in our context, it is a single elliptic solve). Therefore, the cost of applying $\tilde{\boldsymbol{H}}_{\text{misfit}}$ to a vector—and thus the per iteration cost of the randomized algorithm of §V—is dominated by the solution of a pair of linearized forward and adjoint PDEs (explicit expressions for this matrix-vector product will be given for the target problem of inverse wave propagation in §VII).

The remaining component to establish scalability of the low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ is independence of the rank r—and therefore the number of matrix-vector products, and hence PDE solves—from the parameter dimension n. This is the case when $\boldsymbol{H}_{\text{misfit}}$ in (5) is a (discretization of a) compact operator, and when preconditioning by $\boldsymbol{\Gamma}_{\text{prior}}$ does not destroy the spectral decay. This situation is typical for many ill-posed inverse problems, in which the prior is either neutral or of smoothing type (here, we employ a prior that is the inverse of an elliptic operator). Compactness of the data misfit Hessian $\boldsymbol{H}_{\text{misfit}}$ for inverse wave propagation problems has long been observed (e.g., [16]). Recently, we have proven compactness for the inverse wave propagation problem for both continuous and pointwise observation operators for both shape and medium scattering [4], [12]. Specifically, we have shown that the data misfit Hessian is a compact operator. We also quantify the decay of data misfit Hessian eigenvalues in terms of the smoothness of the medium, i.e., the smoother it is, the faster the decay rate. Under some conditions, the rate can be shown to be exponential. That is, the data misfit Hessian can be approximated well with a handful of its dominant eigenvectors and eigenvalues. In conclusion, a low-rank approximation of $\tilde{\boldsymbol{H}}_{\text{misfit}}$ can be made that does not depend on the parameter dimension, and depends only on the information content of the data, filtered through the prior.

Once the r eigenpairs defining the low rank approximation have been computed, estimates of uncertainty can be computed by interrogating $\boldsymbol{\Gamma}_{\text{post}}$ in (9) at a cost of just r inner products (which are negligible) plus elliptic solves representing the action of the square root of the prior $\boldsymbol{\Gamma}_{\text{prior}}^{1/2}$ on a vector (here carried out with algebraic multigrid and therefore scalable). For example, samples can be drawn from the Gaussian defined with a covariance $\boldsymbol{\Gamma}_{\text{post}}$, a row/column of $\boldsymbol{\Gamma}_{\text{post}}$ can be computed, and the action of $\boldsymbol{\Gamma}_{\text{post}}$ in a given direction can be formed, all at cost that is $O(rn)$ for the inner products in addition to the $O(n)$ cost of the multigrid solve. Moreover, the posterior variance field, i.e., the diagonal of $\boldsymbol{\Gamma}_{\text{post}}$, can be found with $O(rn)$ linear algebra plus $O(r)$ multigrid solves.

In summary, we have a method for estimating the posterior covariance—and thus the uncertainty in the solution of the linearized inverse problem—that requires a constant number of PDE solves, dependent only on the information content of the data filtered through the prior (i.e., r), but independent of the number of parameters (n), the number of observations (m), and the number of state variables. Moreover, since the dominant cost of the posterior covariance construction is that of solving forward and adjoint-like PDEs, parallel scalability of the overall uncertainty quantification method follows when the forward PDE solver is scalable (this will be demonstrated for the case of our seismic wave propagation solver in the next section).

VII. Application to global seismic inversion

In recent years, the methodology for scalable parallel solution of forward seismic wave propagation problems on supercomputers by spectral element [17], [18], finite difference [19], finite element [9], and discontinuous Galerkin [20] methods has matured. This motivates our present interest in the seismic inverse problem of determining an earth model from surface observations of seismic waveforms; indeed, we are interested not just in the solution of this inverse problem, but in quantifying the uncertainties in its solution using the method proposed in this paper. In previous sections, our method and underlying algorithms were presented for generic prior and likelihood functions. §VII-A provides explicit expressions for these functions (in infinite dimensions) for the specific seismic inverse problem we address, along with explicit expressions for gradient and Hessian-vector products, which are needed for computing the mean and covariance estimates. The latter expressions involve solutions of forward and adjoint wave propagation PDEs and their linearizations. §VII-B gives an overview of the forward wave equation solver and provides near-full system strong scalability results on the Jaguar supercomputer at ORNL. §VII-C describes the setup of the seismic inverse problem: the configuration of sources and receivers, the

generation of synthetic seismogram observations, the choice of prior and noise covariances, the parametrization of wave speed, and the mesh generation. §VII-C presents results on quantifying uncertainties in the solution of a linearized global seismic inverse problem characterized by one million uncertain parameters. This is the largest—in fact the first—solution of which we are aware of a statistical inverse problem whose forward solver has required a supercomputer, made possible because of the parameter-dimension-independent scaling of our method.

A. Posterior and its derivatives

In this section we give explicit expressions for $V(\boldsymbol{x})$, the negative log of the posterior pdf for the seismic inverse problem we target, along with expressions for its gradient and Hessian-vector product. The expressions are written in strong, infinite-dimensional form, for clarity. The inversion parameter is taken as $c = c(\boldsymbol{x})$, the local acoustic wave speed of the medium. We can write the negative log posterior as

$$\mathcal{V}(c) := \frac{1}{2}\left\|\mathcal{B}\boldsymbol{v}(c) - \boldsymbol{v}^{\text{obs}}\right\|^2_{\Gamma^{-1}_{\text{noise}}} + \frac{1}{2}\left\|c - \bar{c}\right\|^2_{\Gamma^{-1}_{\text{prior}}},$$

where the data misfit (the first term) is a finite dimensional norm due to the pointwise observations in time and space, and the prior term (the second term) is an infinite dimensional norm, with the elliptic prior operator $\Gamma^{-1}_{\text{prior}}$ taken as an anisotropic biharmonic. The wave propagation variables—the velocity vector $\boldsymbol{v}$ and the trace of the strain tensor e (i.e., the dilation) depend on c through the solution of the forward wave propagation equations (written in first-order form):

$$\begin{aligned}
\rho\boldsymbol{v}_t - \nabla(\rho c^2 e) &= \boldsymbol{g} \;\text{ in }\; \Omega\times(0,T),\\
e_t - \nabla\cdot\boldsymbol{v} &= 0 \;\text{ in }\; \Omega\times(0,T),\\
\rho\boldsymbol{v} = \boldsymbol{0}, e &= 0 \;\text{ in }\; \Omega\times\{t=0\},\\
e &= 0 \;\text{ on }\; \partial\Omega\times(0,T).
\end{aligned}$$

Here, ρ and $\boldsymbol{g}$ are known density and seismic source, $\boldsymbol{v}^{\text{obs}}$ are observations at receivers, $\mathcal{B}$ is an observation operator, and Γ_{prior} and Γ_{noise} are the prior and noise covariance operators.

The adjoint approach allows us to write the gradient at a given point c in parameter space as

$$\mathcal{G}(c) := 2\rho c\int_0^T e(\nabla\cdot\boldsymbol{w})\,dt + \Gamma^{-1}_{\text{prior}}(c-\bar{c}),$$

where the adjoint velocity $\boldsymbol{w}$ and adjoint strain dilation d satisfy the *adjoint wave propagation equations*

$$\begin{aligned}
-\rho\boldsymbol{w}_t + \nabla(c^2\rho d) &= -\mathcal{B}^*\Gamma^{-1}_{\text{noise}}(\mathcal{B}\boldsymbol{v} - \boldsymbol{v}^{\text{obs}}) \;\text{ in }\; \Omega\times(0,T),\\
-d_t + \nabla\cdot\boldsymbol{w} &= 0 \;\text{ in }\; \Omega\times(0,T),\\
\rho\boldsymbol{w} = \boldsymbol{0}, d &= 0 \;\text{ in }\; \Omega\times\{t=T\},\\
d &= 0 \;\text{ on }\; \Gamma\times(0,T).
\end{aligned}$$

The adjoint wave equations are reversed in time and have the data misfit as source term, but otherwise resemble the forward wave equations.

Similarly, the action of the Hessian operator in the direction $\tilde{c}$ at a point c is given by

$$\mathcal{H}(c)\tilde{c} := 2\rho\int_0^T ce(\nabla\cdot\tilde{\boldsymbol{w}}) + c\tilde{e}(\nabla\cdot\boldsymbol{w}) + \tilde{c}e(\nabla\cdot\boldsymbol{w})\,dt + \Gamma^{-1}_{\text{prior}}\tilde{c},$$

where $\tilde{\boldsymbol{v}}$ and $\tilde{e}$ satisfy the *incremental forward wave propagation equations*

$$\begin{aligned}
\rho\tilde{\boldsymbol{v}}_t - \nabla(\rho c^2\tilde{e}) &= \nabla(2\rho c\tilde{c}e) \;\text{ in }\; \Omega\times(0,T),\\
e_t - \nabla\cdot\tilde{\boldsymbol{v}} &= 0 \;\text{ in }\; \Omega\times(0,T),\\
\rho\tilde{\boldsymbol{v}} = \boldsymbol{0}, \tilde{e} &= 0 \;\text{ in }\; \Omega\times\{t=0\},\\
\tilde{e} &= 0 \;\text{ on }\; \Gamma\times(0,T).
\end{aligned}$$

On the other hand, $\tilde{\boldsymbol{w}}$ and $\tilde{d}$ satisfy the *incremental adjoint wave propagation equations*

$$\begin{aligned}
-\rho\tilde{\boldsymbol{w}}_t + \nabla(c^2\rho\tilde{d}) &= -\nabla(2\tilde{c}c\rho d) - \mathcal{B}^*\Gamma^{-1}_{\text{noise}}\mathcal{B}\tilde{\boldsymbol{v}} \;\text{ in }\; \Omega\times(0,T),\\
-d_t + \nabla\cdot\tilde{\boldsymbol{w}} &= 0 \;\text{ in }\; \Omega\times(0,T),\\
\rho\tilde{\boldsymbol{w}} = \boldsymbol{0}, \tilde{d} &= 0 \;\text{ in }\; \Omega\times\{t=T\},\\
\tilde{d} &= 0 \;\text{ on }\; \Gamma\times(0,T).
\end{aligned}$$

The incremental forward and incremental adjoint wave equations are seen to be linearized versions of their forward and adjoint counterparts, and thus differ only in the source terms.[1]

Thus, we see that computation of gradients (as needed in the posterior mean approximation) and Hessian actions on vectors (as needed in the posterior covariance approximation) amount to solution of a pair of forward/adjoint wave equations each.

B. Wave propagation solver and its strong scalability

The forward wave equation, and its three variants (adjoint, incremental forward, incremental adjoint) described in the previous section, are solved using a high-order discontinuous Galerkin (dG) method. Details on the forward solver are provided in [11]; here we summarize the salient features:

- discretization that supports h-nonconforming hexahedral elements on a 2:1 balanced forest-of-octrees mesh;
- an element basis that is a tensor product of Lagrange polynomials of arbitrarily high degree based on the Legendre-Gauss-Lobatto (LGL) nodes;
- LGL numerical quadrature, which produces a diagonal mass matrix;
- solution of the Riemann problem at material interfaces (elastic-elastic, elastic-acoustic, acoustic-acoustic);
- mortar-based implementation of flux on 2:1 nonconforming faces;
- time integration by classical four-stage fourth-order Runge Kutta;
- guaranteed consistency, semi-discrete stability, and optimal order convergence for non-conforming meshes [21].

[1]The infinite dimensional expressions for the gradient and Hessian action given above are actually not consistent with the discrete gradient and Hessian-vector product obtained by first discretizing the negative log posterior and wave equation and then differentiating with respect to parameters. Additional jump terms at element interfaces due to the dG discretization appear; in our implementation, we include these terms to insure consistency with discrete counterparts.

To model global seismic wave propagation, we model the earth as a sphere with a radius of 6,371 km, where the speed of acoustic (pressure) waves varies throughout the domain. To generate the finite element mesh, we decompose the earth into 13 warped cubes. The inner core comprises one central cube, surrounded by two layers of six additional cubes. Each cube is the root of an adaptive octree, which can be arbitrarily refined, thus creating a mesh of curved hexahedral elements. The mesh is aligned to the interface between the outer core and the mantle, and several weaker discontinuities between layers, and refined locally to resolve varying seismic wavelengths up to a target frequency. The wave speed $c(\boldsymbol{x})$ is approximated with piecewise trilinear finite elements, and the wave equation variables (velocity and strain) are discretized using high-order (spectral) discontinuous Galerkin finite elements on the same hexahedral mesh. For the distributed storage and adaptation of both the parameter and wave propagation meshes, we use our `p4est` library of fast forest-of-octree algorithms for scalable adaptive mesh refinement, which have been shown to scale to over 220,000 CPU cores and impose minimal computational overhead [20], [22]. The time spent in meshing is insignificant relative to that of numerical solution of the wave equation.

The central difficulty of UQ is its need for repeated solution of the governing PDE model, in our case the wave propagation equations. Conventional sampling methods will take millions of wave propagation solutions (realistically, much more) to explore the posterior distribution for the million-parameter problem we solve in this section. For the frequencies we target, a single wave propagation solve takes a minute on 64K Jaguar cores; conventional sampling methods are thus out of the question. The low-rank Hessian-based method we have presented here, which captures and exploits the local structure of the posterior in the directions informed by the data by computing curvature information based on additional wave equations (adjoint and incremental forward and adjoint), reduces the number of wave propagation solutions by orders of magnitude. Still, thousands of wave equation solves are needed, and we must use all available computing resources. As a result, we insist on excellent strong scalability of our wave equation solver to achieve acceptable time-to-solution. Taken together, the high-order discretization, discontinuous elements, explicit RK scheme, and space filling curve partitioning underlying our forest-of-octree mesh data structure should yield excellent scalability; indeed, we have shown near ideal parallel efficiency in weak scaling on up to 220,000 cores of the Jaguar system at ORNL [11]. Here, we investigate the extreme limits of strong scaling to determine how fine a granularity one can employ in the repeated wave solutions. Table I shows that our wave equation solver exhibits excellent strong scaling over a wide range of core counts. These results are significant, since we are using just third-order elements (higher order creates more work per element, relative to data movement). For the large problem, for example we maintain 71% parallel efficiency in strong scaling from 1024 to 262,144 cores. The largest core count problem has just 62 elements per core.

TABLE I
STRONG SCALING OF THE FORWARD SOLVER

#cores	time [ms]	elem/core	efficiency [%]
256	1630.80	4712	100.0
512	832.46	2356	98.0
1024	411.54	1178	99.1
8192	61.69	148	82.6
65536	11.79	19	54.0
131072	7.09	10	44.9
262144	4.07	5	39.2
1024	5423.86	15817	100.0
4096	1407.81	3955	96.3
8192	712.91	1978	95.1
16384	350.43	989	96.7
32768	211.86	495	80.0
65536	115.37	248	73.5
131072	57.27	124	74.0
262144	29.69	62	71.4

Strong scaling results on ORNL's Jaguar XK6 system for global seismic wave propagation solutions for two problem sizes. We report the time per time step in milliseconds on meshes with 1,206,050 (upper table) and 16,195,864 (lower table) 3rd order discontinuous Galerkin finite elements, corresponding to 694 million and 9.3 billion spatial degrees of freedom, respectively. The elem/core column reports the maximum number of elements owned by any core. For strong scaling from 256 to 262,144 cores, the parallel efficiency is still as high as 39% for the small problem. For the larger problem and a 256-fold increase in problem size, we find a parallel efficiency of 71%. At 262,144 cores, each core owns just 4 or 5 elements for the small problem, and 61 or 62 elements for the larger problem. The larger run sustains a double precision floating point rate of 111 teraflops per second (based on performance counters from the *PAPI* library [23].

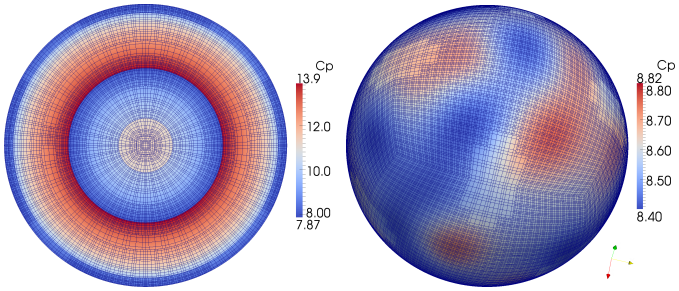

Fig. 2. (Coarser version of) mesh used for the wave propagation simulation and "true" pressure wave speed c in km/s. Left: section through earth model. Right: surface at depth of 222 km showing lateral variations of up to 7%. Wave propagation mesh is tailored to the local seismic wave lengths.

C. Inverse problem solution and its uncertainty

This section presents solution of the statistical inverse problem. First we define the inverse problem setup. Both the prior mean and the initial guess for the iterative solution of the nonlinear least squares optimization problem (2) (to find the MAP point) are derived from the radially symmetric preliminary reference earth model (PREM) [24], which dates to 1981. We take the "true" earth to be given by the more recent S20RTS velocity model (converted from shear to acoustic wave speed anomaly) [25], which superposes lateral wave speed variations on PREM, as seen in Figure 2. Synthetic observations are generated from solution of the wave equation for an S20RTS earth model, with seismic sources at the North pole and at 90° intervals along the equator,

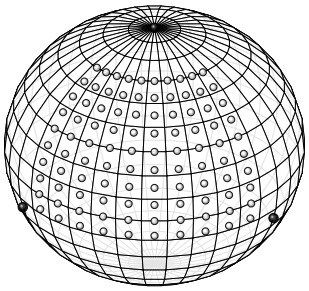

Fig. 3. Location of five simultaneous seismic sources (black spheres; two in back not visible) and 100 receivers (white spheres).

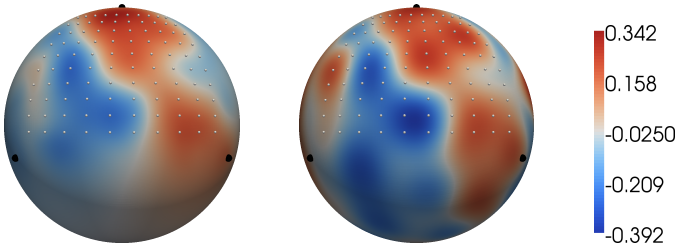

Fig. 4. Comparison of MAP of posterior pdf (left) with the "true" earth model (right) at a depth of 67 km. Source locations are indicated with black spheres and seismic receiver stations are indicated by white spheres.

all of them at a depth of 10 km. All five point sources are taken to occur simultaneously. A total of 100 receivers in the Northern and Eastern hemispheres are distributed along zonal lines at 10° spacing. The source and receiver configuration is illustrated in Figure 3. The observations consist of the first 61 Fourier coefficients of the Fourier-transformed seismogram (time history of ground motion) at each receiver location. The noise distribution for these data is taken as i.i.d. Gaussian with mean zero and a standard deviation of 9.34×10^{-3}.

We use a 3rd-order discontinuous finite elements mesh to resolve seismic wavelengths corresponding to a source with maximum frequency of 0.07 Hz. This requires a mesh with 1,093,784 elements, which leads to 630 million wave propagation spatial unknowns (velocity and strain) for the forward problem, and 1,067,050 unknown wave speed parameters for the statistical inverse problem. The observation time window for the inverse problem is 1,000 seconds, which leads to 2400 discrete time steps. This simulation time is sufficient for the waves to travel about two-thirds of the earth's diameter. A single wave solve takes about one minute on 64K Jaguar cores. As discussed in §VII-A, two wave solves are needed in each gradient or Hessian-vector computation. However, since these expressions combine wave equation solutions in opposite time direction, the work-optimal choice of solving two wave equations requires storage of the entire time history, which is prohibitive. Instead, we use algorithmic checkpointing methods, which cut the necessary storage but increases the number of wave propagation solutions to five per Hessian-vector product (two forward, two incremental forward, and one adjoint solve) [10]. Thus, a single Hessian-vector product takes about 5 minutes on 65K Jaguar cores.

The posterior mean is approximated by solving the nonlinear least squares optimization problem (2) to find the MAP point, using the inexact Gauss Newton-CG method described in §II, initialized with the prior mean (the PREM model), and terminated after 3 orders of magnitude reduction in the gradient, which was achieved after a total of 320 CG iterations (summed across Newton iterations). A comparison of the approximate mean with the "true" earth model (S20RTS) is displayed in Figure 4. The MAP solution is seen to resemble the "true" parameter field well in the Northern hemisphere, which has good receiver coverage.

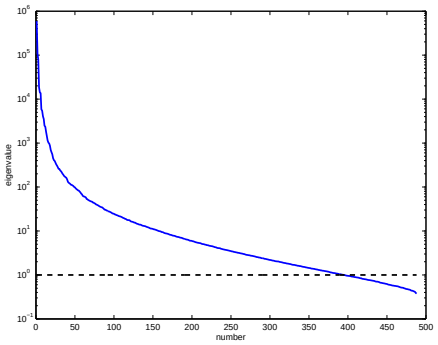

Fig. 5. Logarithmic plot of the spectrum of prior-preconditioned data misfit Hessian.

We approximate the covariance matrix at the MAP point via a low-rank representation employing 488 products of the Hessian matrix with random vectors. The effective problem dimension is thus reduced from 1.07 million to 488, a factor of over 2000 reduction. Figure 5 depicts the first 488 eigenvalues of the million-dimensional parameter field, indicating the rapid decay in information content of the data, a fact that we exploit to make the UQ problem tractable.

The reduction in the variance between prior and posterior due to the information (about the earth model) content of the data—i.e., the diagonal of the second term in (9), the expression for the posterior covariance—is shown in Figure 7. We observe that in the region where sensors are placed (the visible portion of the Northern hemisphere), we get a large reduction in variance due to the data. In regions where there are no sensors, the reduction in variance is substantially less. Additionally, Figure 8 displays the variance reduction on a slice through the equator of the earth, and we again see that the largest variance reduction (depicted in red) is achieved near the surface where the sensors are located, although some reduction is also achieved well into the earth's mantle. Finally, Figure 6 shows samples from the prior and the posterior pdf; the difference between the two sets of samples reflects the information gained from the data in solving the inverse problem. Note the regions of large variability in the posterior samples, which reflect the absence of receivers.

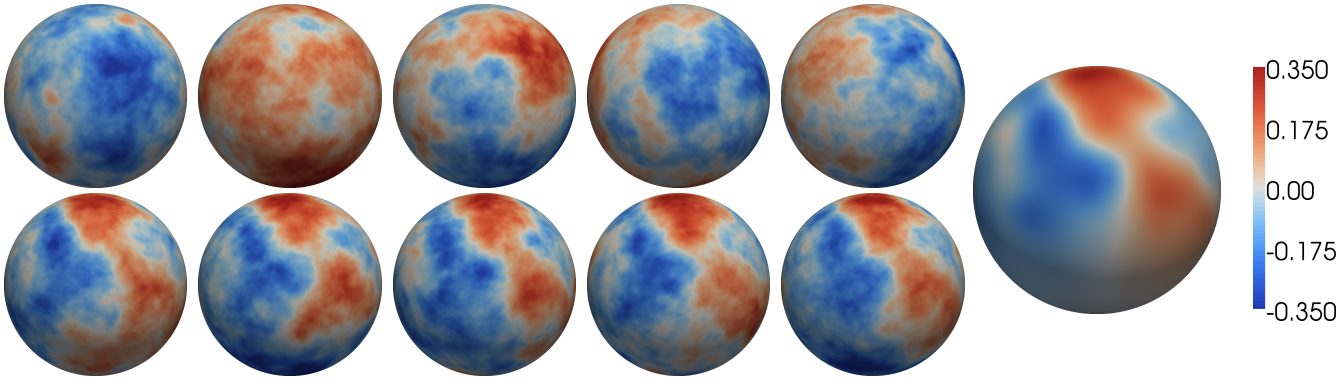

Fig. 6. Samples from the prior (top row) and posterior (bottom row) distributions. The difference between the prior and posterior samples reflects the information (about the earth model) learned from the data. The large scale features of the posterior samples consistently resemble the posterior mean (right). The fine scale features however are not expected to be influenced by the data, and qualitatively resemble the fine scale features of the prior samples. Note the small variability across samples in the Northern hemisphere—reflecting the receiver coverage there—while the Southern hemisphere exhibits large variability in the inferred model, reflecting that uncertainty due to the lack of receivers.

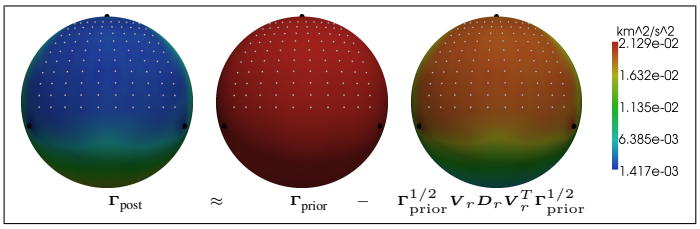

Fig. 7. The left image depicts the pointwise posterior variance field, which is represented as the difference between the original prior variance field (middle), and the reduction in variance due to data (right; see also Figure 8). All variance fields are displayed at a depth of 67km.

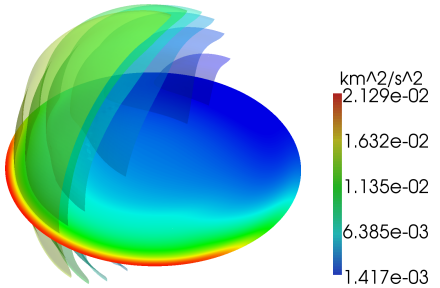

Fig. 8. Data-induced reduction in variance inside the earth. The reduction is shown on a slice through the equator, as well as on isosurfaces in the left hemisphere (compare with Figure 7, which shows reduction on earth's surface). As can be seen, the reduction in variance is greatest on the surface.

VIII. Conclusions

We have addressed UQ for large-scale inverse problems. We adopt the Bayesian inference framework: given observational data and their uncertainty, the governing forward problem and its uncertainty, and a prior probability distribution describing uncertainty in the parameters, find the posterior probability distribution over the parameters. The posterior pdf is a surface in high dimensions, and the standard approach is to sample it via a Markov-chain Monte Carlo (MCMC) method and then compute statistics of the samples. However, the use of conventional MCMC methods becomes intractable for high dimensional parameter spaces and expensive-to-solve forward PDEs, as in our target problem of global seismic inversion.

We have introduced a method that exploits the local structure of the posterior pdf—namely the Hessian matrix of the negative log posterior, which represents the local covariance—to overcome the curse of dimensionality associated with sampling high-dimensional distributions. Unfortunately, straightforward computation of the dense Hessian is prohibitive, requiring as many forward-like solves as there are uncertain parameters. However, the data are typically informative about a low dimensional subspace of the parameter space—that is, the Hessian is sparse with respect to some basis. We have exploited this fact to construct a low rank approximation of the Hessian and its inverse using a matrix-free parallel randomized subspace-detecting algorithm. Overall, our method requires a dimension-independent number of forward PDE solves to approximate the local covariance. Uncertainty quantification for the inverse problem thus reduces to solving a fixed number of forward and adjoint PDEs (which resemble the original forward problem), independent of the problem dimension. The entire process is thus scalable with respect to the forward problem dimension, uncertain parameter dimension, observational data dimension, and number of processor cores. We applied this method to the Bayesian solution of an inverse problem in 3D global seismic wave propagation with one million inversion parameters, for which we observe 3 orders of magnitude dimension reduction, which makes UQ tractable. This is by far the largest UQ problem that has been solved with such a complex governing PDE model.

IX. Acknowledgments

Support for this work was provided by: the U.S. Air Force Office of Scientific Research (AFOSR) Computational Mathematics program under award number FA9550-09-1-0608; the U.S. Department of Energy Office of Science (DOE-SC), Advanced Scientific Computing Research (ASCR), Scientific Discovery through Advanced Computing (SciDAC) program, under award numbers DE-FC02-11ER26052 and DE-FG02-09ER25914, and the Multiscale Mathematics and Optimization

for Complex Systems program under award number DE-FG02-08ER25860; the U.S. DOE National Nuclear Security Administration, Predictive Simulation Academic Alliance Program (PSAAP), under award number DE-FC52-08NA28615; and the U.S. National Science Foundation (NSF) Cyber-enabled Discovery and Innovation (CDI) program under awards CMS-1028889 and OPP-0941678, and the Collaborations in Mathematical Geosciences (CMG) program under award DMS-0724746. Computing time on the Cray XK6 supercomputer (Jaguar) was provided by the Oak Ridge Leadership Computing Facility at Oak Ridge National Laboratory, which is supported by the Office of Science of the Department of Energy under Contract DE-AC05-00OR22725. Computing time on the Texas Advanced Computing Center's Lonestar 4 supercomputer was provided by an allocation from TACC.

REFERENCES

[1] J. T. Oden, R. M. Moser, and O. Ghattas, "Computer predictions with quantified uncertainty, Parts I & II," *SIAM News*, vol. 43, no. 9&10, 2010.

[2] J. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, ser. Applied Mathematical Sciences. New York: Springer-Verlag, 2005, vol. 160.

[3] A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation*. Philadelphia, PA: SIAM, 2005.

[4] T. Bui-Thanh and O. Ghattas, "Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves," *Inverse Problems*, vol. 28, no. 5, p. 055002, 2012.

[5] J. Martin, L. C. Wilcox, C. Burstedde, and O. Ghattas, "A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion," *SIAM Journal on Scientific Computing*, vol. 34, no. 3, pp. A1460–A1487, 2012.

[6] A. M. Stuart, "Inverse problems: A Bayesian perspective," *Acta Numerica*, vol. 19, pp. 451–559, 2010.

[7] T. Bui-Thanh, O. Ghattas, J. Martin, and G. Stadler, "A computational framework for infinite-dimensional Bayesian inverse problems. Part I: The linearized case," 2012, submitted.

[8] V. Akçelik, G. Biros, and O. Ghattas, "Parallel multiscale Gauss-Newton-Krylov methods for inverse wave propagation," in *Proceedings of IEEE/ACM SC2002 Conference*, Baltimore, MD, Nov. 2002, SC2002 Best Technical Paper Award.

[9] V. Akçelik, J. Bielak, G. Biros, I. Epanomeritakis, A. Fernandez, O. Ghattas, E. J. Kim, J. Lopez, D. R. O'Hallaron, T. Tu, and J. Urbanic, "High resolution forward and inverse earthquake modeling on terascale computers," in *SC03: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2003, Gordon Bell Prize for Special Achievement.

[10] I. Epanomeritakis, V. Akçelik, O. Ghattas, and J. Bielak, "A Newton-CG method for large-scale three-dimensional elastic full-waveform seismic inversion," *Inverse Problems*, vol. 24, no. 3, p. 034015 (26pp), 2008.

[11] L. C. Wilcox, G. Stadler, C. Burstedde, and O. Ghattas, "A high-order discontinuous Galerkin method for wave propagation through coupled elastic-acoustic media," *Journal of Computational Physics*, vol. 229, no. 24, pp. 9373–9396, 2010.

[12] T. Bui-Thanh and O. Ghattas, "Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves," *Inverse Problems*, vol. 28, no. 5, p. 055001, 2012.

[13] N. Halko, P. Martinsson, and J. Tropp, "Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions," *SIAM review*, vol. 53, no. 2, pp. 217–288, 2011.

[14] E. Liberty, F. Woolfe, P. Martinsson, V. Rokhlin, and M. Tygert, "Randomized algorithms for the low-rank approximation of matrices," *Proceedings of the National Academy of Sciences*, vol. 104, no. 51, p. 20167, 2007.

[15] H. P. Flath, L. C. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas, "Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations," *SIAM Journal on Scientific Computing*, vol. 33, no. 1, pp. 407–432, 2011.

[16] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering*, 2nd ed., ser. Applied Mathematical Sciences, Vol. 93. Berlin, Heidelberg, New-York, Tokyo: Springer-Verlag, 1998.

[17] D. Komatitsch, S. Tsuboi, C. Ji, and J. Tromp, "A 14.6 billion degrees of freedom, 5 teraflops, 2.5 terabyte earthquake simulation on the Earth Simulator," in *SC03: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2003.

[18] L. Carrington, D. Komatitsch, M. Laurenzano, M. M. Tikir, D. Michéa, N. L. Goff, A. Snavely, and J. Tromp, "High-frequency simulations of global seismic wave propagation using SPECFEM3D GLOBE on 62K processors," in *SC08: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2008.

[19] Y. Cui, K. B. Olsen, T. H. Jordan, K. Lee, J. Zhou, P. Small, D. Roten, G. Ely, D. K. Panda, A. Chourasia, J. Levesque, S. M. Day, and P. Maechling, "Scalable earthquake simulation on petascale supercomputers," in *SC10: Proceedings of the International Conference for High Performance Computing, Networking, Storage, and Analysis*. ACM/IEEE, 2010.

[20] C. Burstedde, O. Ghattas, M. Gurnis, T. Isaac, G. Stadler, T. Warburton, and L. C. Wilcox, "Extreme-scale AMR," in *SC10: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*. ACM/IEEE, 2010, Gordon Bell Prize finalist.

[21] T. Bui-Thanh and O. Ghattas, "Analysis of an *hp*-non-conforming discontinuous Galerkin spectral element method for wave propagation," *SIAM Journal on Numerical Analysis*, vol. 50, no. 3, pp. 1801–1826, 2012.

[22] C. Burstedde, L. C. Wilcox, and O. Ghattas, "`p4est`: Scalable algorithms for parallel adaptive mesh refinement on forests of octrees," *SIAM Journal on Scientific Computing*, vol. 33, no. 3, pp. 1103–1133, 2011.

[23] Performance applications programming interface (PAPI). [Online]. Available: http://icl.cs.utk.edu/papi/

[24] A. M. Dziewonski and D. L. Anderson, "Preliminary reference earth model," *Physics of the Earth and Planetary Interiors*, vol. 25, no. 4, pp. 297–356, 1981.

[25] H. J. van Heijst, J. Ritsema, and J. H. Woodhouse, "Global P and S velocity structure derived from normal mode splitting, surface wave dispersion and body wave travel time data," in *Eos Trans.* AGU, 1999, p. S221.